Name: Seveen Samir Wakim

Id: 2103117

Department: Intelligent systems

Cloud computing Assignment 2 :

Dockerfile and Data Analysis with Popular Books Dataset:

1. **Importing Libraries**: We start by importing the pandas library, which is commonly used for data manipulation and analysis in Python.

```
[3]: import pandas as pd

     # Load the dataset
     data_url = "popular_books_dataset.csv"  # Replace this with the URL to your dataset
```

2. **Loading the Dataset**: We specify the URL of the dataset and use pandas to load it into a DataFrame.

```
[4]: # Load data into a DataFrame
     df = pd.read_csv(data_url)
```

3. **Data Preprocessing**:
    a. **Dropping Irrelevant Columns**: We drop columns that are not required for our analysis, such as IDs and URLs for images.
    b. **Dropping Rows with Missing Values**: We remove any rows that contain missing values.
    c. **Converting Data Types**: We convert the original_publication_year column to numeric data type to ensure consistency.
4. **Filtering the Dataset**: We filter the dataset to focus only on the books from the Harry Potter series.
5. **Further Preprocessing**:
    a. **Removing Duplicate Rows**: We ensure that each book in the Harry Potter series is represented only once by removing duplicate rows based on the title.
    b. **Removing Outliers**: We remove outliers in the dataset based on the ratings count to focus on more representative data.

```
[5]: # Data preprocessing
     # Drop irrelevant columns
     columns_to_drop = ["best_book_id", "work_id", "isbn", "isbn13", "image_url", "small_image_url"]
     df = df.drop(columns=columns_to_drop)

     # Drop rows with missing values
     df = df.dropna()

     # Convert original_publication_year to numeric (if not already)
     df["original_publication_year"] = pd.to_numeric(df["original_publication_year"], errors="coerce")

     # Filter the dataset to focus on the Harry Potter series
     harry_potter_books = df[df["title"].str.contains("Harry Potter")]

     # Further preprocessing
     # Remove duplicate rows based on title
     harry_potter_books = harry_potter_books.drop_duplicates(subset=["title"])

     # Remove outliers based on ratings count
     q_low = harry_potter_books["ratings_count"].quantile(0.01)
     q_high = harry_potter_books["ratings_count"].quantile(0.99)
     harry_potter_books = harry_potter_books[(harry_potter_books["ratings_count"] >= q_low) & (harry_potter_books["ratings_count"] <= q_high)]
```

6. **Analysis**:
    a. **Finding the Most Selling Book**: We find the most selling book within the Harry Potter series based on the highest ratings count.
    b. **Calculating the Average Rating**: We calculate the average rating of the Harry Potter books after preprocessing.

[6]:
```
# Analysis
# Find the most selling book within the Harry Potter series
most_selling_book = harry_potter_books.loc[harry_potter_books["ratings_count"].idxmax()]

# Calculate the average rating of the Harry Potter books
average_rating = harry_potter_books["average_rating"].mean()
```

7. **Printing Results**: We print out the most selling book in the Harry Potter series and the average rating of the Harry Potter books.

[7]:
```
# Print results
print("Most Selling Book in the Harry Potter Series:")
print(most_selling_book[["title", "authors", "average_rating", "ratings_count"]])

print("\nAverage Rating of the Harry Potter Books:", average_rating)
```

```
Most Selling Book in the Harry Potter Series:
title             Harry Potter and the Prisoner of Azkaban (Harr...
authors                     J.K. Rowling, Mary GrandPré, Rufus Beck
average_rating                                                 4.53
ratings_count                                               1832823
Name: 6, dtype: object

Average Rating of the Harry Potter Books: 4.563750000000001
```